Causal inference HW 4

Adeline Shin

12/2/2020

Upload the Data

```
gardasil_df = read.delim("./gardasil.dat.txt", sep = "") %>%
    janitor::clean_names()
```

Question 1

```
# Create subset of data without outcome variable
gardasil_subset = gardasil_df[c(-5)]

describeBy(gardasil_subset, gardasil_subset$practice_type)
```

```
##
##  Descriptive statistics by group
## group: 0
##                vars   n  mean   sd median trimmed  mad min max range  skew
## age               1 515 14.92 2.25     15   14.87 2.97  11  21    10  0.13
## age_group         2 515  0.06 0.25      0    0.00 0.00   0   1     1  3.55
## race              3 515  0.84 0.97      1    0.68 1.48   0   3     3  1.08
## shots             4 515  2.09 0.82      2    2.11 1.48   1   3     2 -0.16
## insurance_type    5 515  1.10 1.15      1    1.00 1.48   0   3     3  0.68
## med_assist        6 515  0.40 0.49      0    0.37 0.00   0   1     1  0.42
## location          7 515  2.74 1.48      4    2.80 0.00   1   4     3 -0.33
## location_type     8 515  0.58 0.49      1    0.60 0.00   0   1     1 -0.33
## practice_type     9 515  0.00 0.00      0    0.00 0.00   0   0     0   NaN
##                kurtosis   se
## age               -0.71 0.10
## age_group         10.62 0.01
## race               0.19 0.04
## shots             -1.50 0.04
## insurance_type    -1.01 0.05
## med_assist        -1.82 0.02
## location          -1.90 0.07
## location_type     -1.90 0.02
## practice_type       NaN 0.00
## ------------------------------------------------------------
## group: 1
##                vars   n  mean   sd median trimmed  mad min max range  skew
## age               1 365 19.46 3.82     19   19.48 4.45  11  26    15 -0.02
## age_group         2 365  0.66 0.47      1    0.70 0.00   0   1     1 -0.69
## race              3 365  1.02 1.19      1    0.91 1.48   0   3     3  0.77
## shots             4 365  2.01 0.81      2    2.01 1.48   1   3     2 -0.01
## insurance_type    5 365  1.85 1.03      1    1.85 0.00   0   3     3  0.13
## med_assist        6 365  0.03 0.18      0    0.00 0.00   0   1     1  5.22
```

```
## location          7 365  1.00 0.00      1    1.00 0.00   1   1     0   NaN
## location_type     8 365  0.00 0.00      0    0.00 0.00   0   0     0   NaN
## practice_type     9 365  1.00 0.00      1    1.00 0.00   1   1     0   NaN
##                kurtosis   se
## age               -0.78 0.20
## age_group         -1.53 0.02
## race              -0.98 0.06
## shots             -1.47 0.04
## insurance_type    -1.74 0.05
## med_assist        25.29 0.01
## location            NaN 0.00
## location_type       NaN 0.00
## practice_type       NaN 0.00
## ------------------------------------------------------------
## group: 2
##                vars   n  mean   sd median trimmed  mad min max range  skew
## age               1 533 21.43 3.33     22   21.63 4.45  11  26    15 -0.46
## age_group         2 533  0.82 0.38      1    0.90 0.00   0   1     1 -1.66
## race              3 533  0.56 0.88      0    0.35 0.00   0   3     3  1.67
## shots             4 533  2.09 0.85      2    2.11 1.48   1   3     2 -0.18
## insurance_type    5 533  1.21 0.78      1    1.14 0.00   0   3     3  1.03
## med_assist        6 533  0.11 0.31      0    0.01 0.00   0   1     1  2.47
## location          7 533  1.99 1.02      2    1.87 1.48   1   4     3  0.67
## location_type     8 533  0.28 0.45      0    0.23 0.00   0   1     1  0.96
## practice_type     9 533  2.00 0.00      2    2.00 0.00   2   2     0   NaN
##                kurtosis   se
## age               -0.70 0.14
## age_group          0.76 0.02
## race               2.01 0.04
## shots             -1.60 0.04
## insurance_type     0.82 0.03
## med_assist         4.13 0.01
## location          -0.73 0.04
## location_type     -1.08 0.02
## practice_type       NaN 0.00
```

Question 2

The question of interest being addressed in this RCT is whether type of practice where Gardasil vaccine is taken affects rates of completion.

i) Treatment and Control Arm

In this study, the treatment arm would consist of those who go to an OB-GYN office to receive their gardasil shots, while the control arm would consist of those who go to a practice to receiver their gardasil shots. Those in the control arm include those under the age of 18 who go to their pediatrician for their gardasil vaccine, as well as adults who visit their general practitioner for the gardasil vaccine.

ii) Eligibility Criteria

```
tableone::CreateTableOne(vars = c("age", "age_group", "race", "shots", "insurance_type", "med_assist",
```

```
##                              Stratified by location
##                               1            2            3
##   n                             798          165           89
##   age (mean (SD))            18.80 (4.14) 20.88 (3.46) 21.92 (3.10)
##   age_group (mean (SD))       0.54 (0.50)  0.76 (0.43)  0.90 (0.30)
##   race (mean (SD))            0.92 (1.15)  0.26 (0.58)  0.81 (0.65)
##   shots (mean (SD))           2.10 (0.83)  2.31 (0.79)  1.74 (0.83)
##   insurance_type (mean (SD))  1.78 (1.03)  1.15 (0.48)  0.73 (0.69)
##   med_assist (mean (SD))      0.04 (0.20)  0.03 (0.17)  0.40 (0.49)
##   location (mean (SD))        1.00 (0.00)  2.00 (0.00)  3.00 (0.00)
##   location_type (mean (SD))   0.00 (0.00)  0.00 (0.00)  1.00 (0.00)
##   practice_type (mean (SD))   1.00 (0.74)  2.00 (0.00)  2.00 (0.00)
##                              Stratified by location
##                               4            p      test
##   n                             361
##   age (mean (SD))            16.09 (3.61) <0.001
##   age_group (mean (SD))       0.20 (0.40) <0.001
##   race (mean (SD))            0.71 (0.83) <0.001
##   shots (mean (SD))           1.97 (0.81) <0.001
##   insurance_type (mean (SD))  0.57 (0.72) <0.001
##   med_assist (mean (SD))      0.56 (0.50) <0.001
##   location (mean (SD))        4.00 (0.00) <0.001
##   location_type (mean (SD))   1.00 (0.00) <0.001
##   practice_type (mean (SD))   0.34 (0.76) <0.001
```

Looking at the descriptive statistics by location, locations 2 and 3 seem to be OB-GYN offices, since they only practice types that correspond to OB-GYN. Furthermore, location 4 does not have any subjects who visit a family practice. Therefore, only location 1 will be included since it is the only one with all three types of practices. This enforces the probabilistic assumption, since it ensures that all subjects have the chance of being a part of either the treatment arm or the control arm.

Question 3

```
# Only including those who are in location 1
gardasil_included = subset(gardasil_subset, location == 1)

# Creating a group for both general practice and pediatrician offices
gardasil_included = gardasil_included %>%
    mutate(practice_type = recode(gardasil_included$practice_type, `0` = 0, `1` = 0, `2` = 1))

# New descriptive statistics
describeBy(gardasil_included, gardasil_included$practice_type)
```

```
##
##  Descriptive statistics by group
## group: 0
##                vars   n  mean   sd median trimmed  mad min max range  skew
## age               1 581 17.76 3.95     17   17.53 4.45  11  26    15  0.44
```

```
## age_group         2 581  0.43 0.50      0    0.41 0.00   0   1     1  0.29
## race              3 581  1.01 1.17      1    0.88 1.48   0   3     3  0.82
## shots             4 581  2.14 0.82      2    2.18 1.48   1   3     2 -0.27
## insurance_type    5 581  1.88 1.06      1    1.92 1.48   0   3     3 -0.03
## med_assist        6 581  0.05 0.22      0    0.00 0.00   0   1     1  3.96
## location          7 581  1.00 0.00      1    1.00 0.00   1   1     0   NaN
## location_type     8 581  0.00 0.00      0    0.00 0.00   0   0     0   NaN
## practice_type     9 581  0.00 0.00      0    0.00 0.00   0   0     0   NaN
##                kurtosis   se
## age               -0.64 0.16
## age_group         -1.92 0.02
## race              -0.87 0.05
## shots             -1.48 0.03
## insurance_type    -1.69 0.04
## med_assist        13.74 0.01
## location            NaN 0.00
## location_type       NaN 0.00
## practice_type       NaN 0.00
## ------------------------------------------------------------
## group: 1
##                vars   n  mean   sd median trimmed  mad min max range  skew
## age               1 217 21.59 3.27     22   21.84 2.97  11  26    15 -0.61
## age_group         2 217  0.85 0.36      1    0.93 0.00   0   1     1 -1.92
## race              3 217  0.69 1.06      0    0.49 0.00   0   3     3  1.38
## shots             4 217  1.98 0.84      2    1.98 1.48   1   3     2  0.03
## insurance_type    5 217  1.52 0.89      1    1.42 0.00   0   3     3  1.01
## med_assist        6 217  0.01 0.10      0    0.00 0.00   0   1     1 10.20
## location          7 217  1.00 0.00      1    1.00 0.00   1   1     0   NaN
## location_type     8 217  0.00 0.00      0    0.00 0.00   0   0     0   NaN
## practice_type     9 217  1.00 0.00      1    1.00 0.00   1   1     0   NaN
##                kurtosis   se
## age               -0.37 0.22
## age_group          1.71 0.02
## race               0.43 0.07
## shots             -1.58 0.06
## insurance_type    -0.86 0.06
## med_assist       102.53 0.01
## location            NaN 0.00
## location_type       NaN 0.00
## practice_type       NaN 0.00
```

Comparing the descriptive statistics created here with those from Question 1, the entirety of group 0 has been left out, since those are the subjects who visited a pediatrics practice for their gardasil vaccine. In addition, there are less subjects in both the OB-GYN and pediatrics groups, since everyone under the age of 18 has been excluded. Therefore, the mean age is a bit higher in both remaining groups. It also looks like the mean number of shots received went down slightly in both groups after excluding those ineligible for the study.

Question 4

```
# Recode practice_type so that 0 = general practice and 1 = OB-GYN office
ps.model = glm(practice_type ~ age + race + shots + insurance_type + med_assist, data = gardasil_includ
```

```r
ps = predict(ps.model, type = "response")

summary(ps.model)
```

```
##
## Call:
## glm(formula = practice_type ~ age + race + shots + insurance_type +
##     med_assist, family = binomial, data = gardasil_included)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6054  -0.7244  -0.4562   0.8903   2.3269
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.75122    0.65489  -7.255 4.02e-13 ***
## age             0.23183    0.02494   9.294  < 2e-16 ***
## race           -0.24996    0.08287  -3.016  0.00256 **
## shots          -0.03635    0.11050  -0.329  0.74218
## insurance_type -0.27373    0.09809  -2.791  0.00526 **
## med_assist     -1.86886    0.78455  -2.382  0.01721 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 933.93  on 797  degrees of freedom
## Residual deviance: 768.35  on 792  degrees of freedom
## AIC: 780.35
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

Question 5

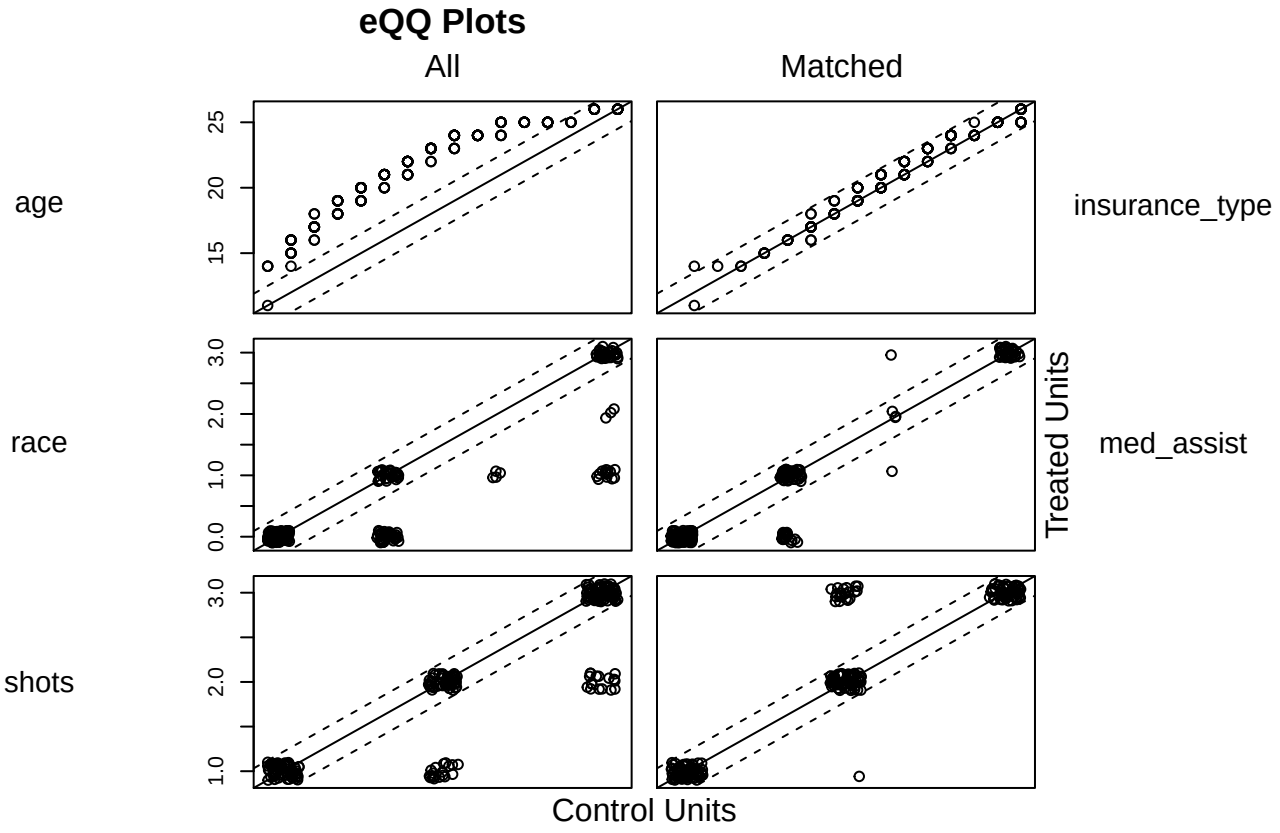
```r
# Nearest neighbor matching with greedy matching
psmatch1 = matchit(practice_type ~ age + race + shots + insurance_type + med_assist, data = gardasil_in

summary(psmatch1, standardize = TRUE)
```

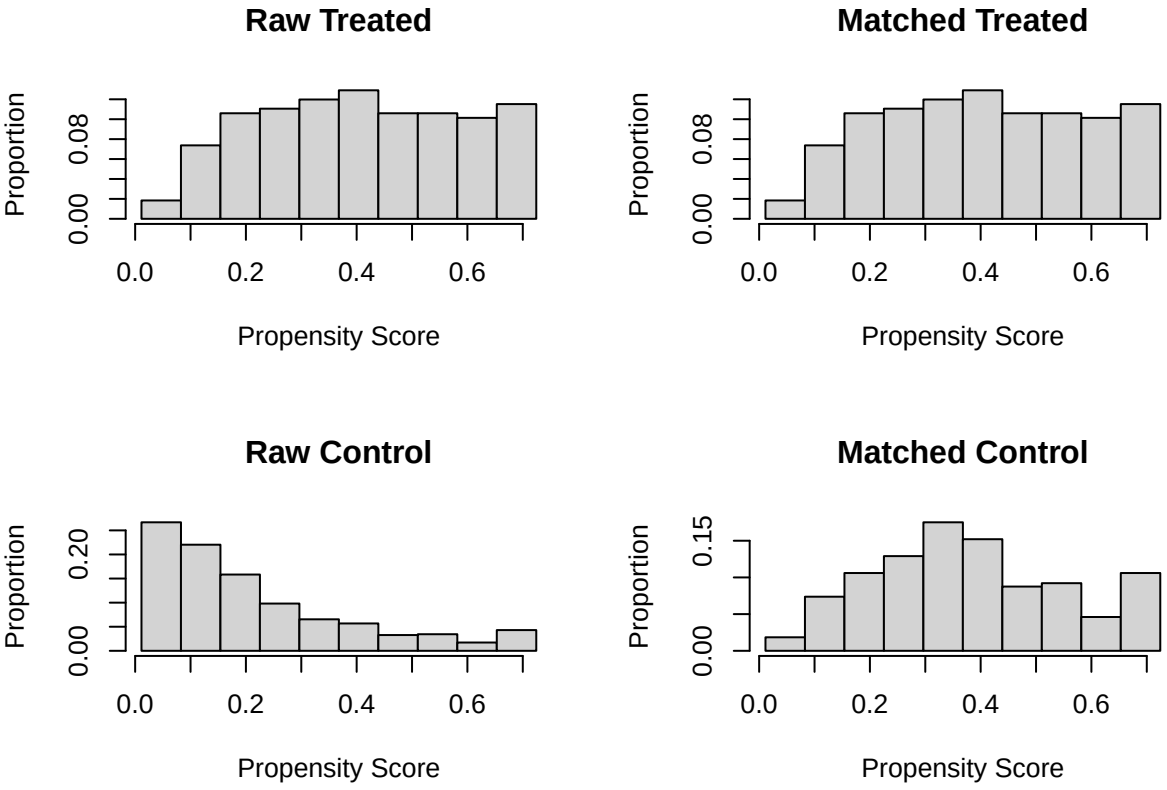
```
##
## Call:
## matchit(formula = practice_type ~ age + race + shots + insurance_type +
##     med_assist, data = gardasil_included, method = "nearest",
##     distance = "logit", discard = "control")
##
## Summary of Balance for All Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance              0.4131        0.2192          1.0638     1.0137
## age                  21.5853       17.7608          1.1708     0.6844
## race                  0.6866        1.0052         -0.2993     0.8265
```

```
## shots                1.9816        2.1446         -0.1944     1.0358
## insurance_type       1.5207        1.8830         -0.4057     0.7068
## med_assist           0.0092        0.0534         -0.4619          .
##                eCDF Mean eCDF Max
## distance          0.2853   0.4774
## age               0.2390   0.4448
## race              0.0796   0.1626
## shots             0.0543   0.0823
## insurance_type    0.1126   0.2112
## med_assist        0.0441   0.0441
##
##
## Summary of Balance for Matched Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance              0.4131        0.3884          0.1355     1.0965
## age                  21.5853       21.2719          0.0959     0.9904
## race                  0.6866        0.7880         -0.0953     1.0270
## shots                 1.9816        1.9032          0.0934     1.1780
## insurance_type        1.5207        1.5161          0.0052     0.9883
## med_assist            0.0092        0.0138         -0.0482          .
##                eCDF Mean eCDF Max Std. Pair Dist.
## distance          0.0205   0.1152          0.1370
## age               0.0253   0.1014          0.3809
## race              0.0253   0.0922          0.7360
## shots             0.0323   0.0876          0.8297
## insurance_type    0.0012   0.0046          0.5523
## med_assist        0.0046   0.0046          0.1447
##
## Percent Balance Improvement:
##                Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                  87.3     -575.2      92.8     75.9
## age                       91.8       97.5      89.4     77.2
## race                      68.2       86.0      68.2     43.3
## shots                     51.9     -365.9      40.6     -6.3
## insurance_type            98.7       96.6      99.0     97.8
## med_assist                89.6          .      89.6     89.6
##
## Sample Sizes:
##           Control Treated
## All           581     217
## Matched       217     217
## Unmatched     261       0
## Discarded     103       0
```

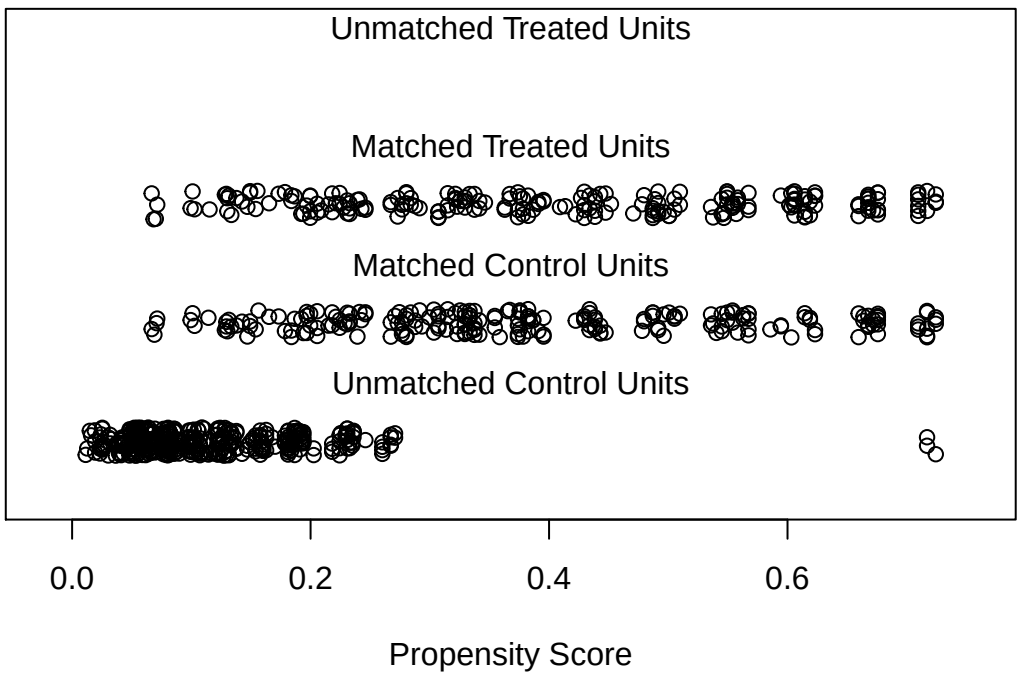
```
plot(psmatch1)
```




```
plot(psmatch1, type="hist")
```



```r
par(mfrow = c(1, 1))
plot(psmatch1, type="jitter", interactive=FALSE)
```



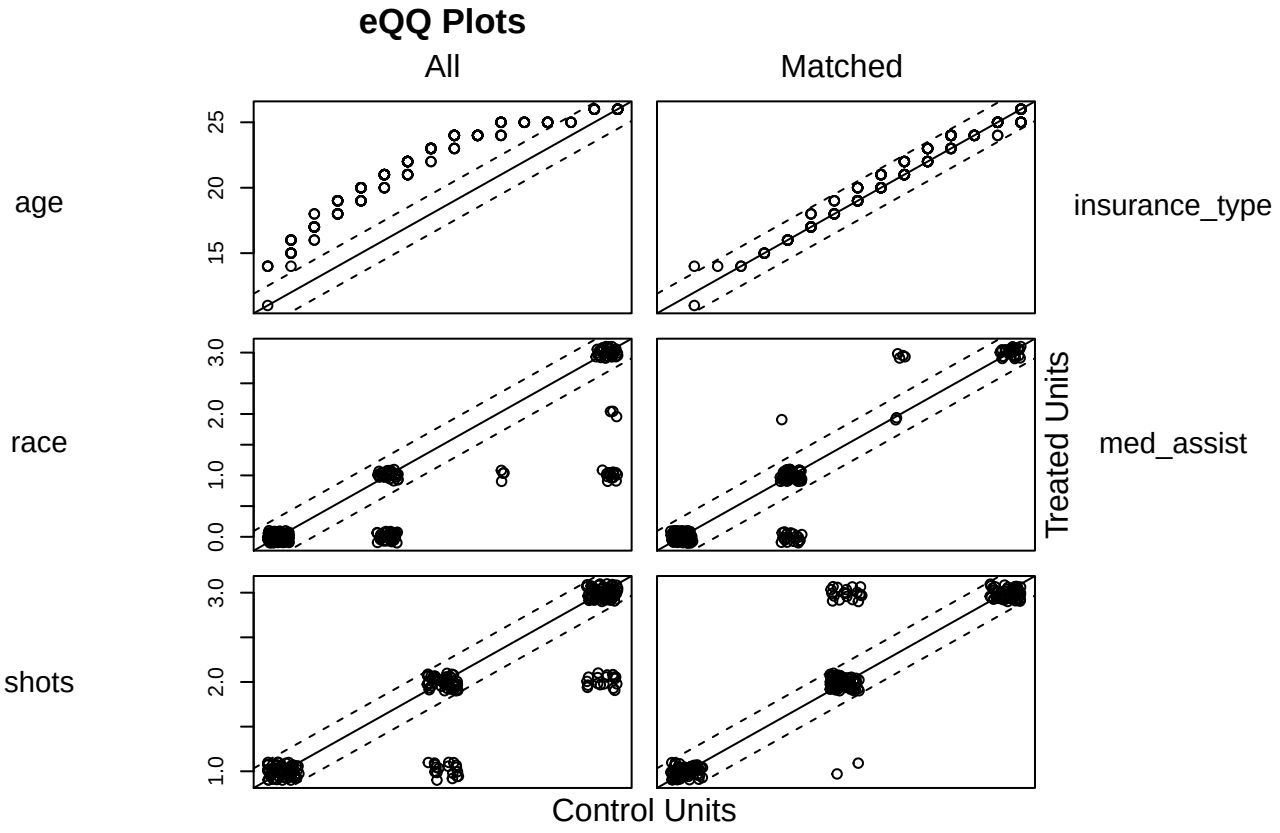

```r
# Nearest neighbor matching with optimal matching instead of greedy
psmatch2 = matchit(practice_type ~ age + race + shots + insurance_type + med_assist, data = gardasil_in

summary(psmatch2, standardize = TRUE)
```

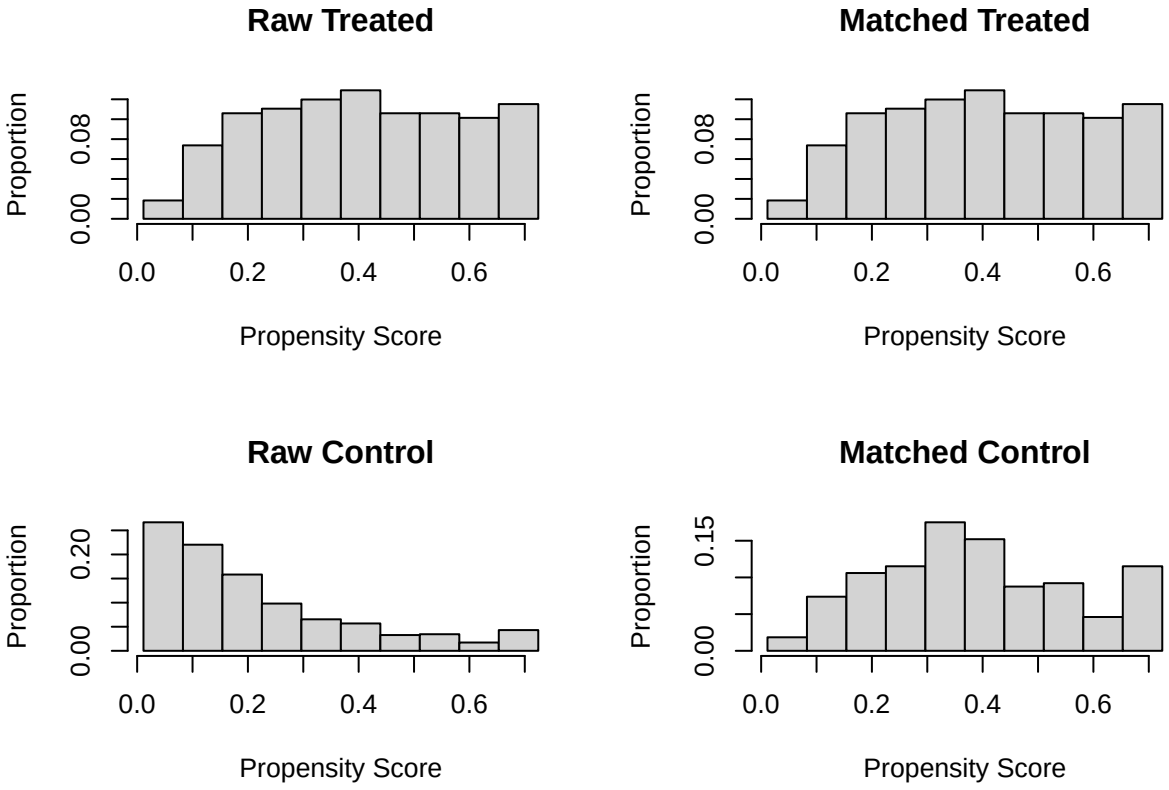
```
##
## Call:
## matchit(formula = practice_type ~ age + race + shots + insurance_type +
##     med_assist, data = gardasil_included, method = "optimal",
##     distance = "logit")
##
## Summary of Balance for All Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance              0.4131        0.2192          1.0638     1.0137
## age                  21.5853       17.7608          1.1708     0.6844
## race                  0.6866        1.0052         -0.2993     0.8265
## shots                 1.9816        2.1446         -0.1944     1.0358
## insurance_type        1.5207        1.8830         -0.4057     0.7068
## med_assist            0.0092        0.0534         -0.4619          .
##                eCDF Mean eCDF Max
## distance           0.2853   0.4774
## age                0.2390   0.4448
## race               0.0796   0.1626
## shots              0.0543   0.0823
## insurance_type     0.1126   0.2112
## med_assist         0.0441   0.0441
```

```
##
##
## Summary of Balance for Matched Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance              0.4131        0.3945          0.1021     1.0507
## age                  21.5853       21.3502          0.0720     0.9573
## race                  0.6866        0.7650         -0.0736     1.0874
## shots                 1.9816        1.9124          0.0824     1.1747
## insurance_type        1.5207        1.5069          0.0155     0.9995
## med_assist            0.0092        0.0138         -0.0482          .
##                eCDF Mean eCDF Max Std. Pair Dist.
## distance          0.0168   0.1014          0.1138
## age               0.0222   0.0876          0.3400
## race              0.0265   0.0922          0.6798
## shots             0.0323   0.0829          0.8077
## insurance_type    0.0035   0.0046          0.4801
## med_assist        0.0046   0.0046          0.1447
##
## Percent Balance Improvement:
##                Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                 90.4     -262.3      94.1     78.8
## age                      93.9       88.5      90.7     80.3
## race                     75.4       56.0      66.7     43.3
## shots                    57.6     -357.8      40.6     -0.7
## insurance_type           96.2       99.9      96.9     97.8
## med_assist               89.6          .      89.6     89.6
##
## Sample Sizes:
##           Control Treated
## All           581     217
## Matched       217     217
## Unmatched     364       0
## Discarded       0       0
```

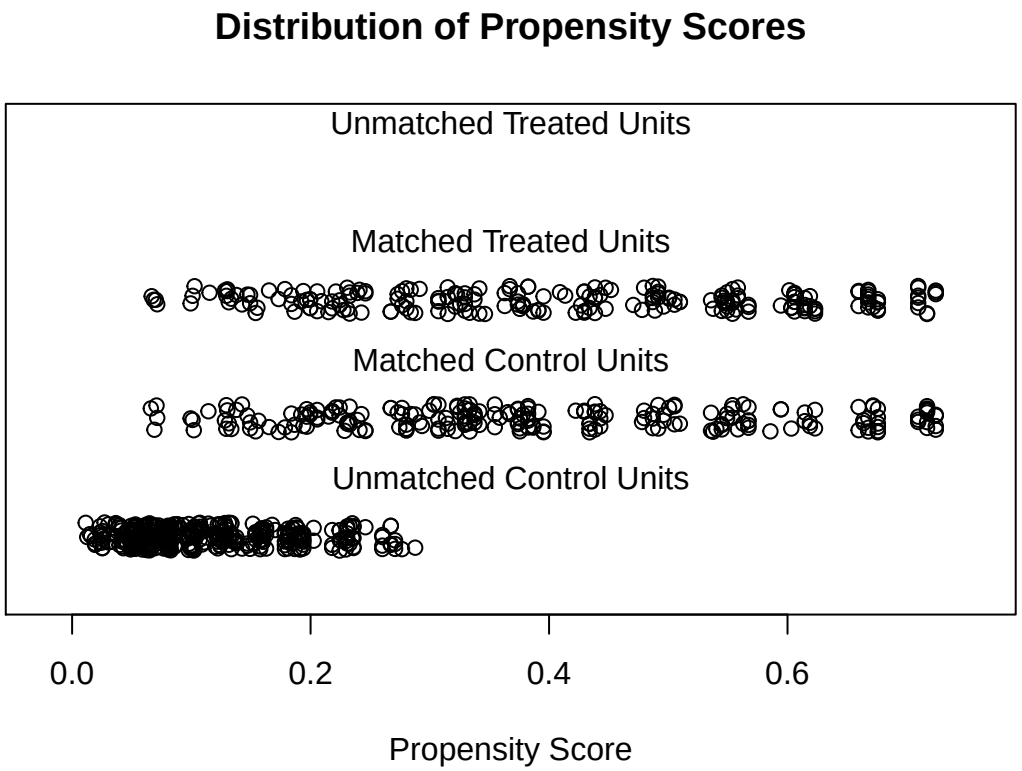
```
plot(psmatch2)
```




```
plot(psmatch2, type="hist")
```



```
par(mfrow = c(1, 1))
plot(psmatch2, type="jitter", interactive=FALSE)
```



The process for matching involved getting rid of the variables location and location type, since they were the same for all subjects after limiting the eligibility to only those in location 1. Both greedy and optimal methods for